D140–D144 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj112

miRBase: microRNA sequences, targets and gene nomenclature

Sam Griffiths-Jones*, Russell J. Grocock, Stijn van Dongen, Alex Bateman and Anton J. Enright

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received September 12, 2005; Revised and Accepted October 18, 2005

ABSTRACT

The miRBase database aims to provide integrated interfaces to comprehensive microRNA sequence data, annotation and predicted gene targets. miRBase takes over functionality from the microRNA Registry and fulfils three main roles: the miRBase Registry acts as an independent arbiter of microRNA gene nomenclature, assigning names prior to publication of novel miRNA sequences. miRBase Sequences is the primary online repository for miRNA sequence data and annotation. miRBase Targets is a comprehensive new database of predicted miRNA target genes. miRBase is available at http://microrna.sanger.ac.uk/.

INTRODUCTION

MicroRNAs (miRNAs) are a class of non-coding RNA gene whose final product is a ∼22 nt functional RNA molecule. They play important roles in the regulation of target genes by binding to complementary regions of messenger transcripts to repress their translation or regulate degradation (1–3). miRNAs have been implicated in cellular roles as diverse as developmental timing in worms, cell death and fat metabolism in flies, haematopoiesis in mammals, and leaf development and floral patterning in plants [reviewed in (4,5)]. Recent reports have suggested that miRNAs may play roles in human cancers (6–8).

The biogenesis of miRNA sequences has been largely elucidated [reviewed in (9)]. The mature miRNA (often designated miR) is processed from a characteristic stem–loop sequence (called a pre-mir), which in turn may be excised from a longer primary transcript (or pri-mir). Only a handful of primary transcripts have been fully described, but evidence suggests that miRNAs are transcribed by RNA polymerase II, and that the transcripts are capped and polyadenylated.

Since the discovery of the founding members of the miRNA class, lin-4 and let-7 in *Caenorhabditis elegans* [reviewed in (10)], over 2000 miRNA sequences have been described in vertebrates, flies, worms and plants, and even in viruses. However, the functions of only a handful of these miRNAs have been experimentally determined. In parallel with novel gene identification efforts, the miRNA community is therefore focused on predicting and validating miRNA gene targets.

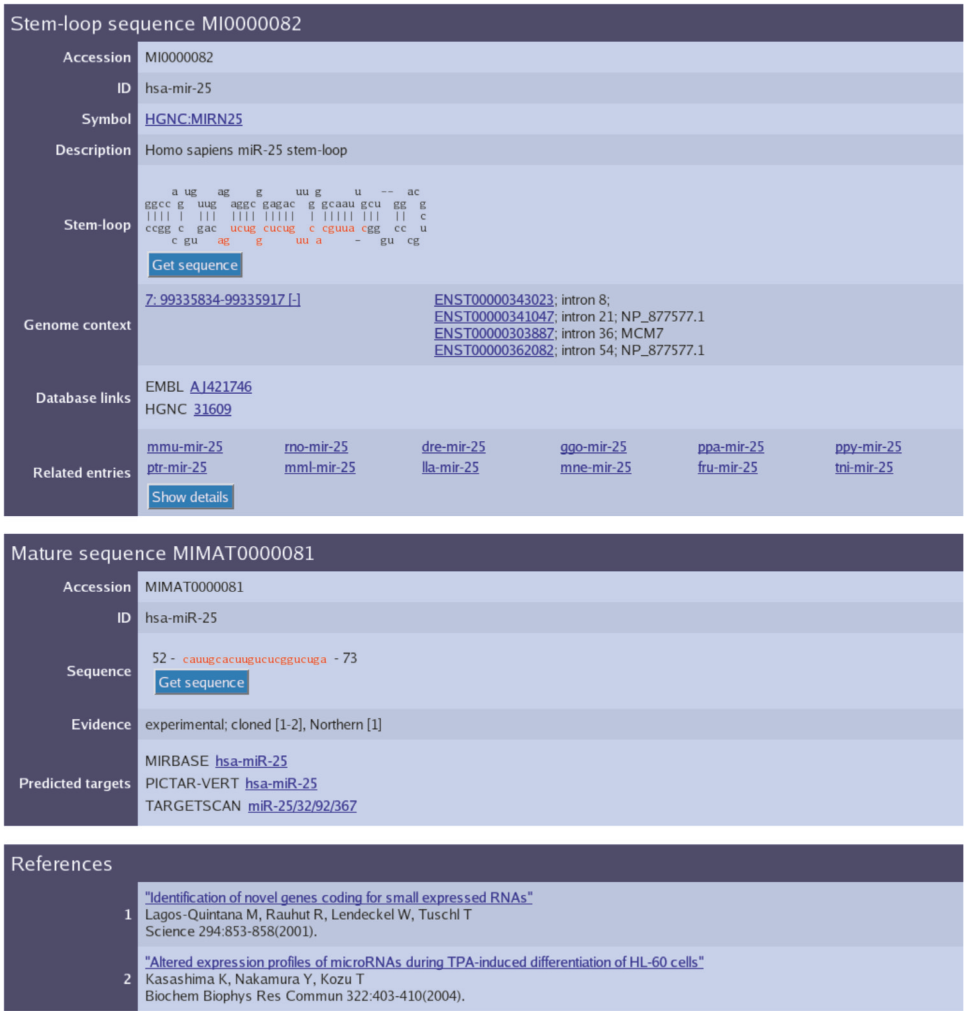
The miRBase database brings together the gene naming and sequence database roles previously fulfilled by the microRNA Registry (11), with the first automated pipeline for predicting miRNA target genes in multiple animal genomes. These three functions are briefly discussed in turn.

miRBase REGISTRY

The rapid growth of the miRNA field has been facilitated by the adoption of a consistent gene naming scheme, which has been applied since the first large-scale miRNA discoveries (12–14). The miRNA Registry (11) has acted as an independent arbiter of gene names, and this function is continued by the miRBase Registry. Names are assigned by the Registry based on guidelines agreed by a number of prominent miRNA researchers and discussed elsewhere (15). In order to minimize the gaps in the naming scheme and to take advantage of the peer review process to assess the validity of submitted miRNAs, names are assigned after a manuscript describing their discovery is accepted for publication. Official gene names should be incorporated into the final version of a manuscript. The nomenclature guidelines require that novel miRNA genes are experimentally verified by cloning or with evidence of expression and processing. Homologous miRNAs from related organisms that are identified by sequence analysis methods may be named without the need for further experimental evidence.

miRNAs are assigned sequential numerical identifiers. The database uses abbreviated 3 or 4 letter prefixes to designate the species, such that identifiers take the form hsa-miR-101 (in *Homo sapiens*). The mature sequences are designated 'miR' in the database, whereas the precursor hairpins are labelled 'mir'. The gene names are intended to convey limited information about functional relationships between mature miRNAs. For example, hsa-miR-101 in human and mmu-miR-101 in mouse

*To whom correspondence should be addressed. Tel: +44 1223 834244; Fax: +44 1223 494919; Email: sgj@sanger.ac.uk



Stem-loop sequence MI0000082

Accession	MI0000082
ID	hsa-mir-25
Symbol	HGNC:MIRN25
Description	Homo sapiens miR-25 stem-loop

Stem-loop

```
    a ug   ag    g     uu g     u   --  ac
ggcc g  uug   aggc gagac  g gcaau gcu  gg  g
|||| |  |||   |||| |||||  | ||||| |||  ||  c
ccgg c  gac   ucug cucug  c cguua cgg  cc  u
    c gu   ag    g     uu a     -   gu  cg
```

Genome context

7: 99335834-99335917 [-]

ENST00000343023; intron 8;
ENST00000341047; intron 21; NP_877577.1
ENST00000303887; intron 36; MCM7
ENST00000362082; intron 54; NP_877577.1

Database links

EMBL AJ421746
HGNC 31609

Related entries

mmu-mir-25	rno-mir-25	dre-mir-25	ggo-mir-25	ppa-mir-25	ppy-mir-25
ptr-mir-25	mml-mir-25	lla-mir-25	mne-mir-25	fru-mir-25	tni-mir-25

Mature sequence MIMAT0000081

Accession	MIMAT0000081
ID	hsa-miR-25
Sequence	52 - cauugcacuugucucggucuga - 73
Evidence	experimental; cloned [1-2], Northern [1]

Predicted targets

MIRBASE hsa-miR-25
PICTAR-VERT hsa-miR-25
TARGETSCAN miR-25/32/92/367

References

1 "Identification of novel genes coding for small expressed RNAs"
Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T
Science 294:853-858(2001).

2 "Altered expression profiles of microRNAs during TPA-induced differentiation of HL-60 cells"
Kasashima K, Nakamura Y, Kozu T
Biochem Biophys Res Commun 322:403-410(2004).

Figure 1. The sequence database entry for hsa-mir-25. The three sections of the page describe the predicted stem–loop hairpin, mature sequences and primary references. The genomic coordinates and contextual information link to the Ensembl database. Each mature miRNA contains an evidence field, and links are provided to predicted target pages.

are orthologous. Paralogous sequences whose mature miRNAs differ at only one or two positions are given lettered suffixes—for example, mmu-miR-10a and mmu-miR-10b in mouse. Distinct hairpin loci that give rise to identical mature miRNAs have numbered suffixes (e.g. dme-mir-281-1 and dme-mir-281-2 in *Drosophila melanogaster*). It should be noted that plant and viral naming schemes differ subtly.

However, miRNA names should not be relied upon to convey complex relationship information. Naming criteria may be subtly redefined over time, and opinion on the degree of conservation of mature sequence required for functional redundancy varies—some recent studies suggest that only the 5′ so-called seed region of the sequence forms a tight duplex with the target mRNA (16). Related hairpin precursor sequences may give rise to mature sequences with only marginal similarity and different miRNA numbers. The naming scheme is also complicated by instances where two different mature miRNA sequences appear to be excised from opposite arms of the same hairpin precursor. Such mature sequences are currently named of the form miR-17-5p (5′ arm) and miR-17-3p (3′ arm). Complex sequence relationships and names are discussed with the submitting author on a case by case basis.

miRBase SEQUENCES

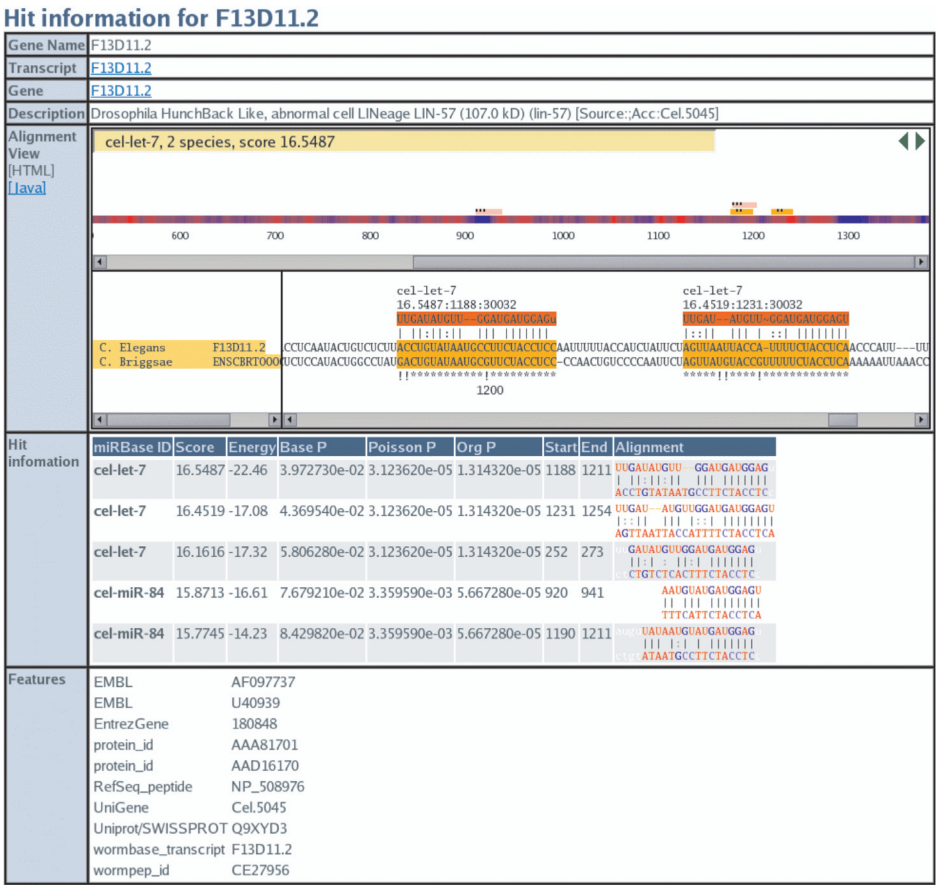
In parallel with the miRNA community's need for a consistent naming scheme, miRNA research and informatics has benefited greatly from a dedicated database of miRNA sequences and annotation. The miRBase Sequence database takes over from the microRNA Registry database as the primary repository for miRNA data. We briefly describe recent growth and database improvements.

Rapid database growth

The miRBase Sequence database contains sequences of all published mature miRNA sequences, together with their predicted source hairpin precursors and annotation relating to their discovery, structure and function. The database has grown rapidly in the past 2 years, from 506 entries representing miRNA hairpin precursors in six species (release 2.0, June 2003) to 2909 entries in 36 species (release 7.0, June 2005).

Stable accessions

miRNA names may change in time to reflect newly discovered relationships between sequences. Stable database accession



Hit information for F13D11.2

Gene Name	F13D11.2
Transcript	F13D11.2
Gene	F13D11.2
Description	Drosophila HunchBack Like, abnormal cell LINeage LIN-57 (107.0 kD) (lin-57) [Source:;Acc:Cel.5045]

Alignment View [HTML] [Java]

cel-let-7, 2 species, score 16.5487

```
                                  cel-let-7                                              cel-let-7
                                  16.5487:1188:30032                                     16.4519:1231:30032
                                  UUGAUAUGUU--GGAUGAUGGAGu                               UUGAU--AUGUU~GGAUGAUGGAGU
                                  | ||:||:||  ||| |||||||                                |::||  ||| | ::| ||||||||
C. Elegans    F13D11.2   CCUCAAUACUGUCUCUUACCUGUAUAAUGCCUUCUACCUCCAAUUUUUACCAUCUAUUCUAGUUAAUUACCA-UUUUCUACCUCAACCCAUU---UU
C. Briggsae   ENSCBRT000UCUCCAUACUGGCCUAUGACUGUAUAAUGCGUUCUACCUCC-CCAACUGUCCCCAAUUCUAGUUAUGUACCGUUUUUCUACCUCAAAAAAUUAAACC
                                  !!***********!**********                               *****!!****!**************
                                                1200
```

Hit infomation

miRBase ID	Score	Energy	Base P	Poisson P	Org P	Start	End	Alignment
cel-let-7	16.5487	-22.46	3.972730e-02	3.123620e-05	1.314320e-05	1188	1211	UUGAUAUGUU--GGAUGAUGGAGu | ||:||:|| ||| ||||||| ACCTGTATAATGCCTTCTACCTCc
cel-let-7	16.4519	-17.08	4.369540e-02	3.123620e-05	1.314320e-05	1231	1254	UUGAU--AUGUUGGAUGAUGGAGU |::|| ||| |::| |||||||| AGTTAATTACCATTTTCTACCTCA
cel-let-7	16.1616	-17.32	5.806280e-02	3.123620e-05	1.314320e-05	252	273	uuGAUAUGUUGGAUGAUGGAGu ||:| : ||:| ||||||| ctCTGTCTCACTTTCTACCTCc
cel-miR-84	15.8713	-16.61	7.679210e-02	3.359590e-03	5.667280e-05	920	941	AAUGUAUGAUGGAGU || ||| |||||||| TTTCATTCTACCTCA
cel-miR-84	15.7745	-14.23	8.429820e-02	3.359590e-03	5.667280e-05	1190	1211	auguUAUAAUGUAUGAUGGAGu ||| |:| | ||||||| ctgtATAATGCCTTCTACCTCc

Features

EMBL	AF097737
EMBL	U40939
EntrezGene	180848
protein_id	AAA81701
protein_id	AAD16170
RefSeq_peptide	NP_508976
UniGene	Cel.5045
Uniprot/SWISSPROT	Q9XYD3
wormbase_transcript	F13D11.2
wormpep_id	CE27956

Figure 2. miRBase Target view page for transcript F13D11.2. The alignment view shows the alignment of miRNA binding sites in orthologous 3′-UTRs. Bits scores, *P*-values, folding energies and alignments are shown for each miRNA match.

numbers are therefore assigned to both hairpin (e.g. MI0000015) and mature (e.g. MIMAT0000029) sequences to enable tracking of sequence entities. A summary of the differences between releases is available. In addition, human and mouse gene symbols are provided in consultation with the Genome Nomenclature Committees (HGNC and MGNC).

Evidence tracking

The database contains miRNAs from two fundamentally different sources. Experimentally verified mature miRNAs are annotated with primary literature references and the experimental method used for discovery. The database also contains sequences that are predicted homologs of miRNAs verified in a related organism. For example, 223 of 313 distinct mature miRNA sequences from human (71%) have experimental evidence in human, while the remainder are clearly identifiable homologs of verified miRNAs from mouse, rat and zebrafish. Homologs are predicted based on sequence similarity and folding characteristics of the precursor hairpin, synteny analysis and conservation of the mature miRNA. The source of every miRNA is clearly annotated on the miRNA entry page (Figure 1) and distributed in the flat file downloads. The miRBase Sequence database does not currently contain predicted miRNAs that are without experimental evidence in any related organism.

Genomic context

For organisms with an assembled genome sequence we provide coordinates of the genomic position of each miRNA sequence on the entry page (Figure 1) and also in GFF format on the FTP site. miRNA genes may be located within other genes, both protein-coding and non-coding (17,18), and the context of the genomic location with respect to Ensembl genes is also annotated (Figure 1). 35% of mammalian miRNA loci overlap annotated genes—over 90% of these are located in introns. In comparison, ∼14% of worm and fly miRNAs are intronic. Distributed annotation system (DAS) sources provide easy access to miRNA genomic locations, and the data are available for viewing within the Ensembl (19) and UCSC browsers (20).

miRBase TARGETS

As focus shifts from miRNA gene identification to functional characterization, miRBase includes not only miRNA sequence data but also information about their genomic targets. The function of a specific miRNA can be thought of as a product of the genes that it regulates. Although large-scale experimental detection of targets is currently difficult, a number of computational techniques exist for the prediction of miRNA targets in mRNA sequences (16,21–27). These methods can be used both to predict potential targets for miRNAs and for the selection of targets for experimental validation. For the most part, computational methods rely on first detecting potential binding sites (with a large degree of complementarity to the miRNA), followed by filtering out those sites that do not appear to be conserved in multiple species. This approach appears to work well, at least for species that have clearly defined orthologs in closely related species (e.g. human, mouse and rat). However, the conservation criterion is poor for those species for which we do not have closely neighbouring genome sequences.

The miRBase Targets database uses a novel fully automated pipeline (which will be described in detail elsewhere) to address some of these issues. All animal miRNA sequences from the miRBase Sequence database are scanned against 3′-untranslated regions (3′-UTRs) predicted from all available species in Ensembl (19) along with *Caenorhabditis briggsae* and *Drosophila pseudoobscura*. The core algorithm assigns *P*-values to individual miRNA–target binding sites, multiple sites in a single UTR, and sites that appear to be conserved in multiple species based on robust statistical models (22). The interface connects each miRNA to a list of predicted gene targets. The detailed target view page (Figure 2) illustrates individual binding sites for one or more miRNAs and their target in an orthologous 3′-UTR alignment. We are in the process of including annotation of experimentally validated miRNA targets.

The miRBase Target database is designed with two main aims: to make available high-quality targets in a timely manner, and to remain as inclusive as possible with respect to the target prediction community. To this end, we provide a core set of predictions that are updated concurrently with the rest of the miRBase system. We also intend to provide a mechanism for viewing and comparing third-party target predictions contributed via DAS. The core predictions are generated in-house using the miRanda algorithm (v3.0) (21). The strengths of miRanda are that it is open source, scalable and incorporates robust statistical models. The provision of a *P*-value for each miRNA–target assignment allows the user to assess the confidence in the prediction. In addition, the method does not assume that the miRNA binding sites *must* be conserved, although in practice the most highly significant *P*-values tend to represent miRNA–target interactions that are conserved across multiple species. As new insights into miRNA–target binding mechanisms and improved prediction algorithms become available, they will be integrated into the system to provide the highest-quality target predictions to the user. In parallel with the miRBase Target pipeline, miRNA sequence entries also provide links to third-party target prediction websites (Figure 1).

AVAILABILITY

The miRBase database is freely available to all for online searching at http://microrna.sanger.ac.uk/. Sequences and annotation are also available for download from the FTP site in a number of formats, including FASTA format sequences and relational database dumps for easy upload to a MySQL or other database. Queries, feedback and data submissions and revisions are welcome by email to microrna@sanger.ac.uk.

ACKNOWLEDGEMENTS

We thank Mhairi Marshall and John Tate for website design, and are grateful to David Bartel and Tom Tuschl for ongoing nomenclature discussion. We also thank Michel Weber for assistance in providing data for viewing in the UCSC genome browser, Marc Rehmsmeier for discussion of *P*-value statistics and Antonio Giraldez for experimental work on target verification. Work at the Sanger Institute is supported by the Wellcome Trust. Funding to pay the Open Access publication charges for this article was provided by the Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. Filipowicz,W., Jaskiewicz,L., Kolb,F.A. and Pillai,R.S. (2005) Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr. Opin. Struct. Biol.*, **15**, 331–341.
3. Sontheimer,E.J. and Carthew,R.W. (2005) Silence from within: endogenous siRNAs and miRNAs. *Cell*, **122**, 9–12.
4. Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.
5. Kidner,C.A. and Martienssen,R.A. (2005) The developmental role of microRNA in plants. *Curr. Opin. Plant Biol.*, **8**, 38–44.
6. He,L., Thomson,J.M., Hemann,M.T., Hernando-Monge,E., Mu,D., Goodson,S., Powers,S., Cordon-Cardo,C., Lowe,S.W., Hannon,G.J. and Hammond,S.M. (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828–833.
7. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
8. O'Donnell,K.A., Wentzel,E.A., Zeller,K.I., Dang,C.V. and Mendell,J.T. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839–843.
9. Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nature Rev. Mol. Cell Biol.*, **6**, 376–385.
10. Pasquinelli,A.E. and Ruvkun,G. (2002) Control of developmental timing by microRNAs and their targets. *Annu. Rev. Cell Dev. Biol.*, **18**, 495–513.
11. Griffiths-Jones,S. (2004) The microRNA Registry. *Nucleic Acids Res.*, **32**, D109–D111.
12. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
13. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, **294**, 858–862.
14. Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, **294**, 862–864.
15. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S., Marshall,M. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
16. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
17. Rodriguez,A., Griffiths-Jones,S., Ashurst,J.L. and Bradley,A. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.
18. Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.
19. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
20. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.

21. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biol*., **5**, R1.
22. Rehmsmeier,M., Steffen,P., Hochsmann,M. and Giegerich,R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
23. Stark,A., Brennecke,J., Russell,R.B. and Cohen,S.M. (2003) Identification of *Drosophila* microRNA targets. *PLoS Biol*, **1**, E60.
24. Rajewsky,N. and Socci,N.D. (2004) Computational identification of microRNA targets. *Dev Biol*., **267**, 529–535.
25. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
26. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA–target recognition. *PLoS Biol*., **3**, e85.
27. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. and Rajewsky,N. (2005) Combinatorial microRNA target predictions. *Nature Genet*., **37**, 495–500.